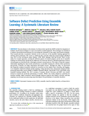
Machine Learning Classifiers A Brief Primer

Abdul Ahad Abro

University of Sindh Journal of Information and Communication Technology (USJICT)



University of Sindh Journal of Information and Communication Technology
(USJICT)
Volume 5, Issue 2, July 2021

ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

Machine Learning Classifiers: A Brief Primer

Abdul Ahad Abro[1], Abdullah Ayub Khan[2], Mir Sajjad Hussain Talpur , Idrissa Kayijuka[4], Erkan Yaşar[1]

[1]Department of Computer Engineering, Ege University, Izmir, Turkey.
[2]Department of Computer Science, Benazir Bhutto Shaheed University Lyari, Karachi.
[3]Information Technology Centre, Sindh Agriculture University, Tandojam.
[4]Department of Applied Status, University of Rwanda, Rwanda.

abdulahadabro1@gmail.com, abdullah.ayub@bbsul.edu.pk, mirsajjadhussain@sau.edu.pk, kayijukai@gmail.com, erkan.yasar@ege.edu.tr

Abstract: Machine learning is a prominent and an intensively studied field in the artificial intelligence area which assists to enhance the performance of classification. In this paper, the main idea is to provide the classification and comparative analysis of data mining algorithms. To support this idea, six supervised machine learning (ML) algorithms, C4.5 (J48), K-Nearest Neighbor (KNN), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), and One Rule (OneR) along with the five UCI Datasets of ML Repository, are being applied that demonstrates the robustness and effectiveness of numerous approaches. Whereas, for analytical procedures, significant parameters have been considered: Accuracy, Area Under Curve (AUC), precision, recall, and F-measure values. Hence, the primary objective of this study is to obtain binary classification and efficiency by conducting the performance evaluation. We present experimental results that demonstrate the effectiveness of our approach to well-known competitive approaches.

Keywords: Artificial Intelligence; Machine Learning; Data Mining; Classification; Knowledge Discovery in Databases.

I. Introduction

Data mining involves the adoption of sophisticated data analysis tools to discover the relation and valid patterns in large datasets [1][2]. Abundant theoretical and empirical studies are published, providing the advantages of the combination paradigm over the individual classifier models [3][4][5]. In recent times, ML is being widely used in a variety of industries, for instance, remote sensing, image classification and pattern recognition.

These tools include interdisciplinary research areas which are arithmetical algorithms, statistical models, ML methods, intelligent information systems, etc [6].

J48 is a simple C4.5 decision tree [7]. The classification process is modelled by applying the binary tree. It is the successor of the ID3 algorithm. In [8], it is an efficient evaluation model and recursively select the attribute with the maximum information gain rate as the test attribute and finally generate the satisfactory outcomes.

KNN is one of the top data mining algorithms; it is significant to extend KNN classifiers sensitive to costs for imbalanced data classification applications. The major weakness of the lazy learning (KNN classification) algorithm is often sensitive to the noisy data and the irrelevant or disturbing attributes. It is an efficient feature selection methods for data mining [9].

LR and Neural Networks were systematically ranked among the best models. It yields magnificent performance as (ML) models to predict the risk of major chronic diseases with low incidence. ML is superior to conventional regression for disease prediction modeling, where the number of incident disease cases is low [10].

The NB classifier aggregates with the Bayes paradigm in decision rules like the hypothesis, which is the possible result. NB learning framework for large-scale computational efficiency and multi-domain platform classification [11].

Support vector machines (SVMs) are powerful and flexible algorithms and can handle multiple continuous and categorical variables. Furthermore, the overall outcomes with comparisons are given, indicating significant non-linear mitigation with BER reductions. The SVM multi-classifier based on the in-phase and quadrature components is relatively optimal, considering the calculation and storage [12].

OneR [13], a simple classification algorithm that generates a one-level decision tree. It handles the missing values, numeric attributes mentioning flexibility and create one rule for each attribute in the training data, then selects the rule with the minimum error rate as its one rule.

This paper is structured in several sections. In section II, the literature survey is briefly described. The proposed methodology adopted for performing different experiments is in Section III. Section IV, states experimental analysis, datasets detail and performance evaluation. Lastly, the conclusion is drawn based on outcomes and future work is suggested in Section V.

II. L ITERATURE REVIEW

Research-based on C4.5, KNN, LR, NB, SVM and OneR, classification has recently witnessed a surge of research efforts [14]. In this paper, we use the binary classification of supervised learning. Classification aims to accurately forecast the target class for each case in the data. The model builds the training process, a classification algorithm discoveries association between the values of the predictors and the values of the target [15]. Different classification algorithms practice different procedures for discovery associations. These associations are model, which can function to a different dataset in which the class is unidentified [16].

In C4.5 [17], application of the C-C4.5 algorithm on noisy data is more robust to noise than the one of the C4.5 algorithm. The performance of the C-C4.5 algorithm depends on the different parameters. The results obtained C-C4.5 trees with high values obtain the best results according to average accuracy.

In [18], KNN is the slowest classification technique because the classification time is directly related to the number of data. When the data size is more prominent, more extensive distance calculation must be performed and this makes KNN extremely slow. It uses the number of nearest neighbors "k" as one of the parameters in classifying an object and the value of k influences the classifier [19].

In [20], investigates the properties of LR and NB and makes a comparison between them. The hybrid LR-NB model construction method that follows the strategy of balancing the tradeoff between model bias and model variance, to minimize the sum of errors. The hybrid model offers an improvement over pure LR in terms of training time by optimizing fewer parameters in the LR part.

In [21], NB is the most popular data mining algorithms. Empirical results indicate that the selective NB demonstrates superior classification performance while retaining the simplicity and flexibility at the same time.

In [22], incorporating imprecise prior knowledge and sophisticated machine learning SVM-based algorithms has been proposed. It uses the duality representation in the framework of the minimax strategy of decision making, which permits us to get simple extensions of SVMs, including additional constraints for optimization variables.

In [13], OneR is a simple classification algorithm that generates a one-level decision tree. It is also able to handle missing values and numeric attributes showing flexibility despite the simplicity. The OneR algorithm creates one rule for each attribute in the training data, then selects the rule with the minimum error rate.

III. PROPOSED METHODOLOGY

This section presents an overview of the proposed method, which describes the pre-processing stage of data and classification algorithms used in this study.

A. *Proposed System*

The proposed system is given in Figure. 1. It consists of numerous phases: datasets, base learners and comparative analysis of results. Besides, the generalization performance of the system, 10-fold cross-validation is used for all classifier learners and datasets.

B. *Data Pre-processing*

The values of ranges in the data from different machine learning datasets may be high. In this case, certain features can significantly or negatively affect algorithms for classification accuracy. Therefore, data values are normalized to [0,1] range using min-max normalization technique [23].

C. *Classification of Algorithms*

In this study, base learners, including C4.5(J48), KNN, LR, NB, SVM and OneR, are employed.

There are numerous phases of method related to datasets and classifiers focused on ML. In this work, six ML classifiers, along with five datasets, are experienced for binary classification.

C4.5 is performing best among all the algorithms such as Naïve Bayes, Random Forest and Multilayer Perceptron. C4.5 is showing the best classification accuracy as compared to NB, Random Forest and Multilayer Perceptron while applying attribute selection [7].

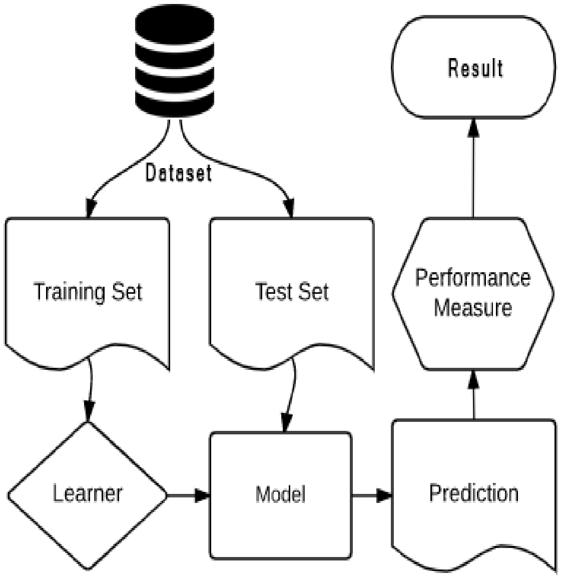
The [24], DPeak clustering algorithm, is not applicable for large scale datasets. In this scenario, the FastDPeak is proposed. Its density is replaced by kNN-density, which is computed by a fast kNN algorithm such as a cover tree, yielding huge improvement for density computations. Experimental results show that FastDPeak is effective and outperforms as compared to other ML algorithms.

LR classifier is another method borrowed by ML from the field of statistics. It is a statistical model and used when the dependent variable is categorical. NB is a probabilistic ML model. It requires linear parameters in the number of functions of the variables and highly scalable [25].

In [26], naive Bayes and random forest had to overlap the performance and both ML methods outperformed multiple logistic regression. These ML methods (multiple logistic regression, naive Bayes, and random forest) were able to predict survival at a population level. It is better to choose the best method for each moment, but all methods would have resulted in similar improvements in practice.

In [27], an algorithm is proposed to optimize the feature subsets of samples, and then add the parameters of the support vector machine to optimize the classification. The experimental results show that the algorithm has a good effect on the classification of adequate instant messaging information of the Internet of things big data and has a good effect and practical application value.

OneR classification algorithm that generates one rule for each predictor in the data. In [28], a procedure for ML is a straightforward one that proves surprisingly useful on the standard datasets commonly used for evaluation. It takes as input a set of examples, each with several attributes and a class like other learning methods. The OneR algorithm selects the most informative single attribute and bases the rule on this attribute alone. However, the result is not satisfactory with continuous-valued attributes and handling the missing values.



.Figure.1. The Proposed Layout

IV. EXPERIMENT AND ANALYSIS

In these subsections, we describe and present the experimental process, evaluation measures and experimental results.

A. Experimental Process

In the experimental process, datasets have been used from the UCI Machine Learning Repository [29].

All experiments are performed on a total of 6 ML classifiers by using WEKA (Waikato Environment for Knowledge Analysis) ML toolkit and JAVA programming language. We utilized default parameter values for all classifiers in WEKA [30].

We carry out 10-fold cross-validation to all datasets to yield reliable results. The 10-fold cross-validation is imposed on the original dataset randomly partitioned into ten equally sized sets, one of which is used as test validation, while the remaining sets are used for training operations. The process is repeated ten times and calculated the averages of the results.

Dataset characteristics are evaluated concerning the attributes and the number of instances. These datasets are typically used to solve machine learning related problems. There are various numerical attributes, instances and class descriptions illustrated in Table I. The datasets are selected from the UCI Machine Learning Repository according to their distinct parameters, which are being utilized for binary classification problems.

TABLE I
DATASETS DETAIL

Datasets	Instances	Attributes	Classes
Adult	48842	14	2
Breast Cancer	286	9	2
Car Evaluation	1728	6	4
Iris	150	4	3
Yeast	1484	8	10

In this work, different supervised ML approaches have been carried out along with the datasets, which are considered suitable for the classification. However, the performance metrics are calculated according to binary classification problems based on the confusion matrix.

B. Assessment of Measures

This section describes the five performance evaluation measures of the proposed method, consisting of accuracy, AUC, precision, recall and F-measure.

Accuracy reflects how close an agreed number is to a measurement. It is specified further in Eq.1.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(1)}$$

In equation 1, TN, FN, FP and TP show the number of True Negatives, False Negatives, False Positives and True Positives.

AUC represents the area under the ROC Curve. AUC calculates the whole two-dimensional area beneath the whole ROC curve from (0,0) to (1,1).

Precision is a positive analytical value [15]. Precision defines how reliable measurements are, although they are farther from the accepted value.

The equation of precision is shown in Eq.2.

$$Precision = \frac{TP}{TP + FP} \quad \textbf{(2)}$$

The recall is the hit rate [15]. The recall is the reverse of precision; it calculates false negatives against true positives. The equation is illustrated in Eq. 3.

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(3)}$$

F-measure can be defined as the weighted average [14] [15], of precision and recall. This rating considers both false positives and false negatives. The equation is illustrated in Eq. 4.

$$F - \text{measure} = \frac{2}{1/\text{precision} + 1/\text{recall}} \qquad \textbf{(4)}$$

These criteria are adjusted proportionally in the data by the reference class prevalence in the weighting operation.

C. *Experimental Results*

There are several algorithms for classification of which the most well-known and widely applicable dataset.

Tables II-VI for all datasets present accuracy, AUC, precision, recall and F-measurement values of machine learning algorithms. In Table II-VI, high Acc, AUC, Precision, Recall and F-measure are shown in Bold, while the greyed shows insufficient results.

To sum up, Tables II-VI, has been designed in terms of different specifications according to the multiple datasets relating to the numerous approaches to machine learning. In Table II, LR has better outcomes, which provides 85.6988% Acc in comparison to others. Probably, in Table III, J48 indicates 75.5245% Acc adequate consequences. Similarly, in Table IV, the SVM presents 93.7500% Acc effective results. Likewise, in Table V, the LR illustrates the 96.1000% Acc productive outcomes. However, in the end, LR shows a 58.6253% Acc result in Table VI.

Moreover, it is analyzed that LR in adult dataset Table II, provides positive findings. Likely, J48 in the breast cancer dataset concerning Table III, indicates the progressive result.

Similarly, Table IV, SVM presents effective results in the car evaluation dataset. Likewise, in Table V, the iris dataset LR provides a more accurate outcome. Finally, LR indicates adequate consequences in Table VI, yeast dataset.

TABLE II
VALUES FOR ADULT DATASET

Adult					
Methods	**Acc (%)**	**AUC**	**Precision**	**Recall**	**F-Measure**
J48	85.1705	0.875	0.845	0.852	0.845
KNN	79.8718	0.722	0.798	0.799	0.798
LR	**85.6988**	0.911	0.851	0.857	0.852
NB	82.3779	0.902	0.846	0.824	0.831
SVM	85.4367	0.764	0.848	0.854	0.848
OneR	79.6569	0.592	0.797	0.797	0.747

TABLE III
VALUES FOR BREAST CANCER DATASET

Breast Cancer					
Methods	**Acc (%)**	**AUC**	**Precision**	**Recall**	**F-Measure**
J48	**75.5245**	0.584	0.752	0.755	0.713
KNN	72.3776	0.628	0.699	0.724	0.697
LR	68.8811	0.646	0.668	0.689	0.675
NB	71.6783	0.701	0.704	0.717	0.708
SVM	69.5804	0.590	0.671	0.696	0.677
OneR	65.7343	0.542	0.624	0.657	0.635

TABLE IV
VALUES FOR CAR EVALUATİON DATASET

Car Evaluation					
Methods	**Acc (%)**	**AUC**	**Precision**	**Recall**	**F-Measure**
J48	92.3611	0.976	0.924	0.924	0.924
KNN	93.5185	0.997	0.940	0.935	0.925
LR	93.1134	0.990	0.932	0.931	0.931
NB	85.5324	0.976	0.852	0.855	0.847
SVM	**93.7500**	0.953	0.939	0.938	0.938
OneR	70.0231	0.500	0.700	0.700	0.824

TABLE V
VALUES FOR IRIS DATASET

Iris					
Methods	**Acc (%)**	**AUC**	**Precision**	**Recall**	**F-Measure**
J48	96.0000	0.968	0.960	0.960	0.960
KNN	95.3333	0.966	0.953	0.953	0.953
LR	**96.1000**	0.981	0.960	0.960	0.960
NB	96.0000	0.994	0.960	0.960	0.960
SVM	96.0000	0.978	0.962	0.960	0.960
OneR	92.0000	0.940	0.920	0.920	0.920

TABLE VI
VALUES FOR YEAST DATASET

Yeast					
Methods	**Acc (%)**	**AUC**	**Precision**	**Recall**	**F-Measure**
J48	55.9299	0.733	0.549	0.559	0.553
KNN	52.2911	0.685	0.524	0.523	0.522
LR	**58.6253**	0.825	0.585	0.586	0.577
NB	57.6146	0.816	0.585	0.576	0.566
SVM	57.0755	0.781	0.478	0.571	0.596
OneR	40.027.0	0.585	0.404	0.400	0.517

In Figure. 2-6, indicate the effects of enhanced classification of performance evaluation on datasets given in charts.

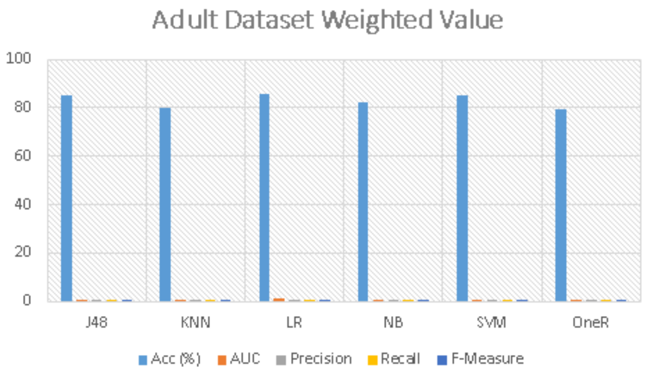


Figure.2. The chart is showing the effects of Adult dataset.

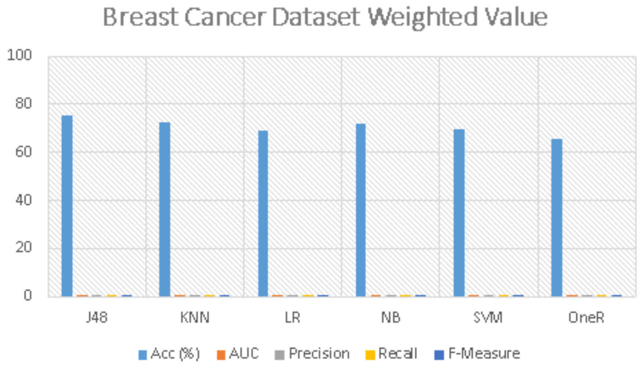


Figure.3. The chart is showing the effects of Breast Cancer dataset

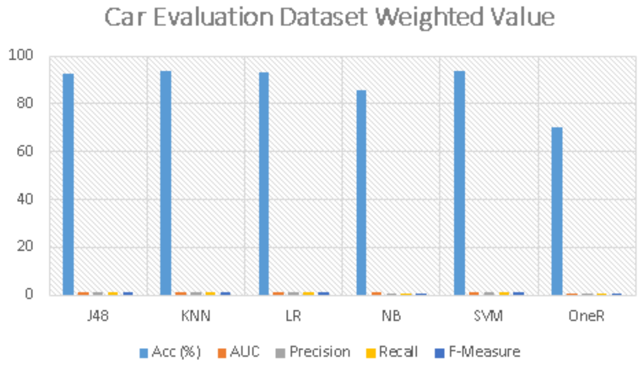


Figure.4. The chart is showing the effects of Car Evaluation dataset.

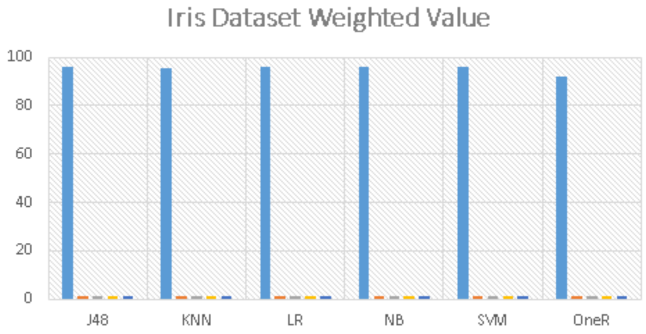


Figure.5. The chart is showing the effects of Iris dataset.

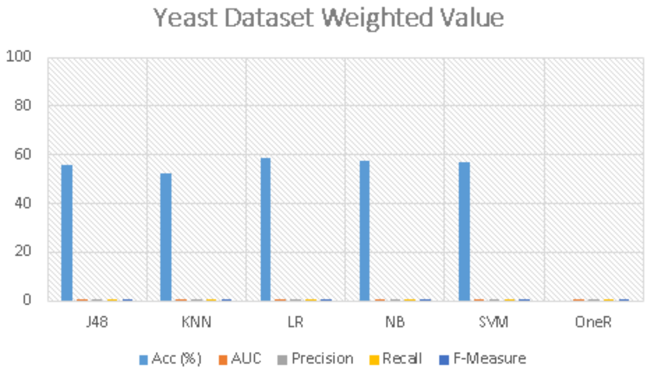


Figure.6. The chart is showing the effects of Yeast dataset.

V. CONCLUSION AND FUTURE WORK

In this paper, we have examined the execution of supervised ML algorithms to classify multiple datasets, namely C4.5 (J48), K-Nearest Neighbor (KNN), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM) and One Rule (OneR). The efficiency of algorithms is further classified in terms of recall/sensitivity, precision, accuracy and F -score. The sensitivity and specificity of the same algorithm can be severely affected by a retrospective study, analyzed varying sizes of training and test sets.

This work can be extended to other data mining techniques like clustering and association.

In the future, we plan to reform our study of classification models by introducing the hybrid framework of intelligent machine learning system will use to an extensive collection of real-life datasets.

REFERENCES

[1] R. Accorsi, R. Manzini, P. Pascarella, M. Patella, and S. Sassi, “Data Mining and Machine Learning for Condition-based Maintenance,” Procedia Manuf., vol. 11, no. June, pp. 1153–1161, 2017.

[2] Y. Shao, Y. Liu, X. Ye, and S. Zhang, “A machine learning based global simulation data mining approach for efficient design changes,” Adv. Eng. Softw., vol. 124, no. July, pp. 22–41, 2018.

[3] E. Hüllermeier, “Fuzzy sets in machine learning and data mining,” Appl. Soft Comput. J., vol. 11, no. 2, pp. 1493–1505, 2011.

[4] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research,” Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017.

[5] M. Shafiq, Z. Tian, A. K. Bashir, A. Jolfaei, and X. Yu, “Data mining and machine learning methods for sustainable smart cities traffic classification: A survey,” Sustain. Cities Soc., vol. 60, no. March, p. 102177, 2020.

[6] S. Deepajothi and S. Selvarajan, “A Comparative Study of Classification Techniques On Adult Data Set 1,” Int. J. Eng. Res. Technol., vol. 1, no. 8, pp. 1–8, 2012.

[7] D. Bansal, R. Chhikara, K. Khanna, and P. Gupta, “Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia,” Procedia Comput. Sci., vol. 132, pp. 1497–1502, 2018.

[8] X. Wang, C. Zhou, and X. Xu, “Application of C4.5 decision tree for scholarship evaluations,” Procedia Comput. Sci., vol. 151, no. 2018, pp. 179–184, 2019.

[9] S. Zhang, “Cost-sensitive KNN classification,” Neurocomputing, vol. 391, pp. 234–242, 2020.

[10] S. Nusinovici et al., “Logistic regression was as good as machine learning for predicting major chronic diseases,” J. Clin. Epidemiol., vol. 122, pp. 56–69, 2020.

[11] F. Xu, Z. Pan, and R. Xia, “E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework,” Inf. Process. Manag., no. February, p. 102221, 2020.

[12] C. Wang et al., “QAM classification methods by SVM machine learning for improved optical interconnection,” Opt. Commun., vol. 444, no. March, pp. 1–8, 2019.

[13] V. S. Parsania, N. N. Jani, and N. H. Bhalodiya, “Applying

Naïve bayes , BayesNet , PART , JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis," Int. J. Darshan Inst. Eng. Res. Emerg. Technol., vol. 3, no. 1, pp. 1–6, 2014.

[14] A. A. Abro, M. A. Yimer, and Z. Bhatti, "Identifying the Machine Learning Techniques for Classification of Target Datasets," Sukkur IBA J. Comput. Math. Sci., vol. 4, no. 1, 2020.

[15] A. A. Abro, E. Taşci, and A. Uğur, "A Stacking-based Ensemble Learning Method for Outlier Detection," Balk. J. Electr. Comput. Eng., vol. 8, no. 2, pp. 181–185, 2020.

[16] Abro, A. A., Soomro, S., Alansari, Z., Belgaum, M. R., & Khakwani, A. B. K. (2016). Secure Network in Business-to-Business application by using Access Control List (ACL) and Service Level Agreement (SLA). arXiv preprint arXiv:1612.07685.

[17] C. J. Mantas, J. Abellán, and J. G. Castellano, "Analysis of Credal-C4.5 for classification in noisy domains," Expert Syst. Appl., vol. 61, pp. 314–326, 2016.

[18] A. Ashari, "Performance Comparison between Naïve Bayes , Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool," vol. 4, no. 11, pp. 33–39, 2013.

[19] M. Bilal, H. Israr, M. Shahid, and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naı ¨ ve Bayesian , Decision Tree and KNN classification techniques," J. King Saud Univ. - Comput. Inf. Sci., vol. 28, no. 3, pp. 330–344, 2016.

[20] Y. Tan and P. P. Shenoy, "A bias-variance based heuristic for constructing a hybrid logistic regression-naïve Bayes model for classification," Int. J. Approx. Reason., vol. 117, pp. 15–28, 2020.

[21] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," Knowledge-Based Syst., vol. 192, p. 105361, 2020.

[22] L. V. Utkin, "An imprecise extension of SVM-based machine learning models," Neurocomputing, vol. 331, pp. 18–32, 2019.

[23] T. Classification and B. K. Singh, "Investigations on Impact of Feature Normalization Techniques on Investigations on Impact of Feature Normalization Techniques on Classifier ' s Performance in Breast Tumor Classification," no. April 2015, pp. 10–15, 2017.

[24] Y. Chen et al., "Fast density peak clustering for large scale data based on kNN," Knowledge-Based Syst., vol. 187, p. 104824, 2020.

[25] P. Tsangaratos and I. Ilia, "Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size," Catena, vol. 145, pp. 164–179, 2016.

[26] E. M. M. van der Heide, R. F. Veerkamp, M. L. van Pelt, C. Kamphuis, I. Athanasiadis, and B. J. Ducro, "Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle," J. Dairy Sci., vol. 102, no. 10, pp. 9409–9421, 2019.

[27] Y. Chen, "Mining of instant messaging data in the Internet of Things based on support vector machine," Comput. Commun., vol. 154, no. March, pp. 278–287, 2020.

[28] C. G. Nevill-Manning, G. Holmes, and I. H. Witten, "The development of Holte's 1R classifier," Proc. - 1995 2nd New Zeal. Int. Two-Stream Conf. Artif. Neural Networks Expert Syst. ANNES 1995, no. January, pp. 239–242, 1995.

[29] UCI Machine Learning Repository, 2018, https://archive.ics.uci.edu/ml/index.php.

[30] T. A. Engel, A. S. Charão, M. Kirsch-Pinheiro, and L. A. Steffenel, "Performance improvement of data mining in weka through GPU acceleration," Procedia Comput. Sci., vol. 32, pp. 93–100, 2014.